

# Quantifying discrepancies in opinion spectra from online and offline networks

Deokjae Lee<sup>1</sup>, Kyu S. Hahn<sup>2</sup>, Soon-Hyung Yook<sup>3</sup>, Juyong Park<sup>4,\*</sup>

**1 Center for Complex Systems Studies and CTP, Department of Physics and Astronomy, Seoul National University, Korea**

**2 Department of Communication, Seoul National University, Korea**

**3 Department of Physics, Kyung Hee University, Korea**

**4 Graduate School of Culture Technology, Korea Advanced Institute of Science & Technology, Korea**

\* juyongp@kaist.ac.kr

## Abstract

Online social media such as Twitter are widely used for mining public opinions and sentiments on various issues and topics. The sheer volume of the data generated and the eager adoption by the online-savvy public are helping to raise the profile of online media as a convenient source of news and public opinions on social and political issues as well. Due to the uncontrollable biases in the population who heavily use the media, however, it is often difficult to measure how accurately the online sphere reflects the offline world at large, undermining the usefulness of online media. One way of identifying and overcoming the online-offline discrepancies is to apply a common analytical and modeling framework to comparable data sets from online and offline sources and cross-analyzing the patterns found therein. In this paper we study the political spectra constructed from Twitter and from legislators' voting records as an example to demonstrate the potential limits of online media as the source for accurate public opinion mining.

## Introduction

The proliferation of online social media services such as Twitter is widely recognized as signifying a revolution in how we utilize information and understand the way our society communicates. It also presents an opportunity open and accessible to any interested party to utilize the massive data accrued from such services for understanding the world as well as ourselves [1, 2]. Among many attempts to harness the potential of online social media, a prominent one is to mine the opinions and sentiments of the public for a variety of purposes, either academic or commercial [3–6]. The use of social media for such purposes poses a potentially critical problem, however: Since online social media users constitute an uncontrolled, thus likely unfair, sample of the population unlike in well-designed opinion polls, it is unclear to what extent the online sphere accurately represents the offline world. User profiles collected by online social media and made available to researchers frequently fall short of giving a sufficiently accurate demographic information about the users, and the traditional polling techniques using questionnaires have to be used to augment them [7]. The ideal case would be that the demographics of online users were identical to that of the offline public at large, guaranteeing the agreement of the opinions mined from the online sphere and with traditional polls. A less ideal, but a manageable case would be when the online demographics are precisely known, allowing us to employ straightforward calibrations to find accurate results. But for a limited number of cases the demographic differences between online service users and non-users are not precisely known [8], and given the breadth of topics discussed online and the

difficulty of accurately estimating the number of users it does not seem realistic that such a state can ever be reached.

To those looking to mine for information from social media as a proxy for conducting polling, the potential disparity between the online and the offline world due to the unknowable demographics raises serious doubts on the usefulness of online social media, particularly where the offline reality is essential such as politics. In such cases the lack of a systematic method by which to compare the online and the offline would make finding relevant, useful information from online data infeasible. When researchers study such problems, they could circumvent this issue by restricting the scope of their study to the online sphere so that by design the real-world public at large is of little concern. This may be appropriate when the issue being studied is pertinent only to the online world; for instance, when the issue affects the online sphere only, or when one declares that their only interest is in the opinions of the online sphere. While certain scientific discoveries can still be made this way since the data themselves are nonetheless novel and interesting, the inability to draw conclusions regarding many issues pertinent to the offline potentially cripples the potential of online social media.

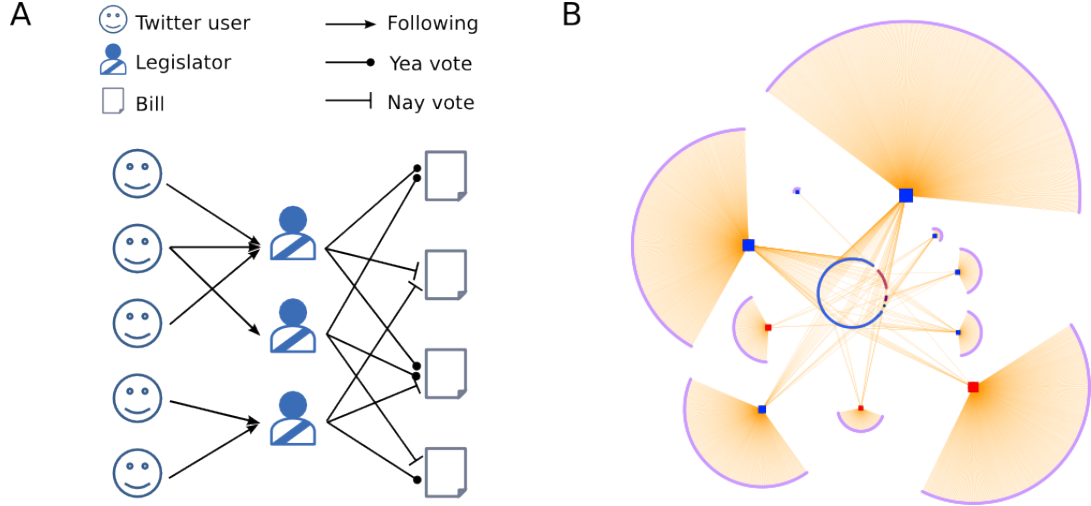
In order to overcome such a problem, it is important to identify the discrepancies between the online and the offline quantitatively. One way of achieving that is to apply a common analytical and modeling framework to comparable data sets representing the online and the offline so that we can perform a cross-analysis of the patterns found therein. In this paper we demonstrate the process using as an example the landscape of political spectra (signifying the political leanings or positions) constructed from distinct online and offline data of the US and South Korea, two nations with modern representative democracies. Specifically we study the political spectra of the legislators in each country, constructed in two ways. First, we construct a spectrum from the Twitter followership network, which we take as representing the online political landscape. There is some direct evidence that connection on social networks are biased towards those with similar political viewpoints. In a well-known study, Adamic and Glance showed a clear partisan divide on the blogosphere [9]. Regarding Twitter, Parmalee and Bichard showed that over 70% of those espousing strong conservative or liberal ideology completely avoided following political leaders who challenge their beliefs [10]. In the case of Korea (one of the two countries studied in this work), Chang found that both politicians and ordinary Twitter users exhibit a systematic relationship between their political choices and their position in the Twitter network [11]. Second, we construct a spectrum from the legislators' roll call (voting) records, which we take as representing the offline political landscape that more closely reflects the true public one, since the legislators' voting is likely to be influenced by the offline public. The cross-examination of the two spectra should then reveal the differences between the online and the offline.

## Results

### Data

We utilized the following four data sets: (i) The Twitter followership network of the members of the 111th U.S. Senate; (ii) the roll call of the Senators on legislative bills; (iii) The Twitter followership network of the members of the 18th Korean National Assembly; and (iv) the roll call of the Assembly members between the years 2008 to 2010 (covering three quarters of the term). The Twitter data are current as of June 2011, and Twitter users who followed two or more legislators were included in the data. The data contain all legislators who owned an official Twitter account at the time of data collection and maintained their seat during the entire term considered. This resulted in 67 senators (out of 100), 698 roll calls, and 139 806 Twitter users for the US, and 194 assembly members (out of 299), 1 119 roll calls, and 124 341 Twitter users for Korea. Both countries have two-party systems, the Republican Party (GOP) and the Democratic Party (DEM) that account for 99% of the US Senate, and the Grand National Party (GNP) and the Korean Democratic Party (DP) that account for 86% of the Korean National Assembly. [12]

The data were modeled as bipartite networks. See Figure 1 for the schematic illustrations of the data.



**Figure 1.** (A) The schematic illustrations for the data employed in our study. The legislators are included in two distinct bipartite networks. On the left is the legislator-Twitter user network, and on the right the legislator-legislative bill network. Of the two types of edges – a ‘nay’ vote and a ‘yea’ vote – we consider the ‘nay’, since they are believed to carry more information in determining the legislators’ political spectra (see the Method section for more detail). The structure of each bipartite network can reveal differences in political positions of the legislators, which is the origin of the online-offline discrepancy. Here, for example, the upper two legislators occupy similar positions that are different from that of the lower one since their follower sets are disjoint. The two groups’ voting patterns may show less clear differences. In this study we use Multidimensional Scaling (MDS) and Kendall’s ranking correlation coefficient to quantify the spectra and their discrepancies. (B) A sample of the US senator-Twitter followership network consisting of ten legislators and their Twitter followers. The red squares are Republican (GOP) senators, and the blue squares are the Democratic (DEM) senators. All other nodes are Twitter users.

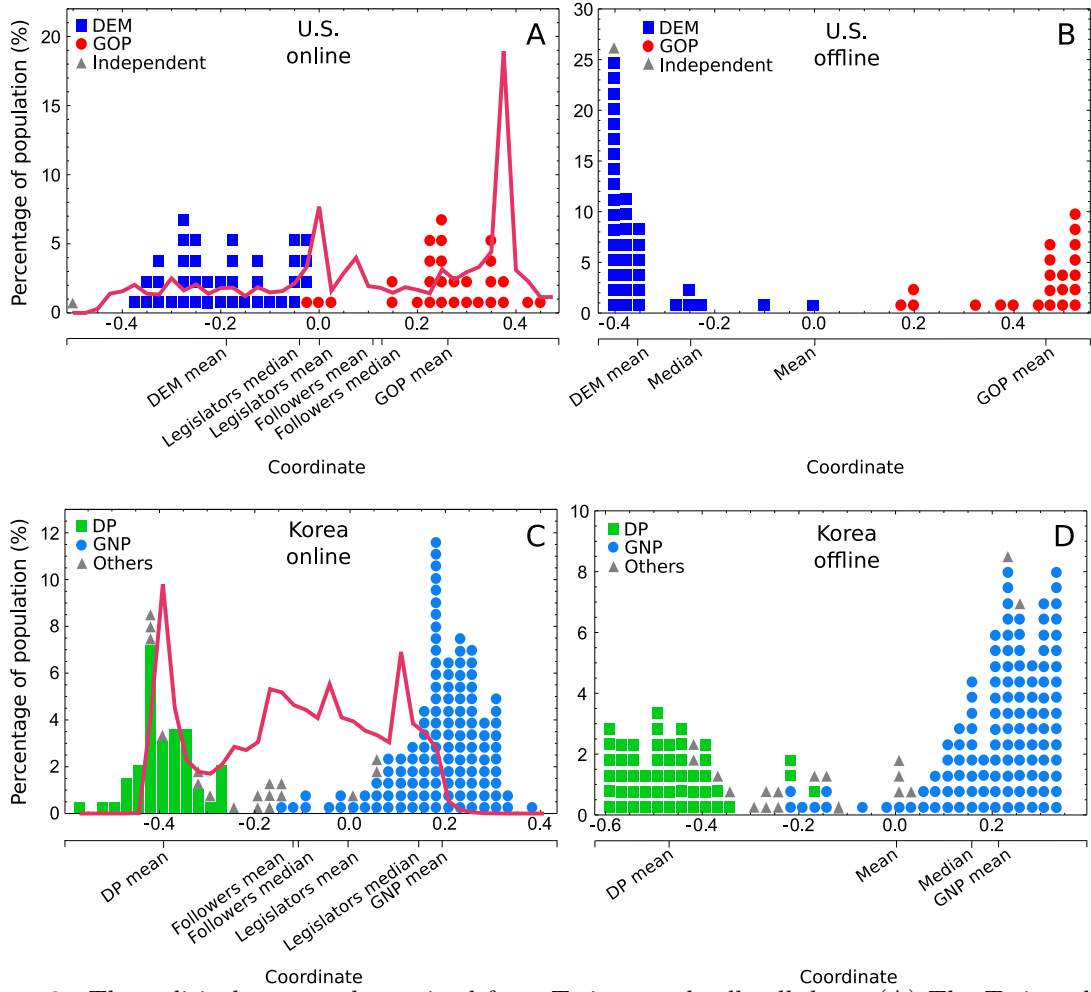
## Online and offline political spectra

The data introduced above can be modeled as bipartite networks on which the political spectra are constructed using matrix decomposition methods [13,14]. In this study we use Classical Multidimensional Scaling (CMDS) [15,16]. The method requires a dissimilarity function that act as the distance between two legislators. We use the Kulczynski dissimilarity [17]. In essence, CMDS regards the dissimilarity values as the Euclidean distances between legislators in 1-dimensional space and determines the coordinates of the legislators that most closely matches the given distances. We then interpret the coordinates as the political positions of the legislators. Once the spectra of the legislators are determined this way using the Twitter network, we can construct the spectra of the Twitter users by averaging the positions of the legislators that each user follows (see the Methods section for more details).

We consider the offline, roll call-based spectra as revealing the true political positions of the legislators as they results from their actual actions of voting. Furthermore, we accept the roll call-based spectra as more accurately reflecting true political positions of the offline public and the society, as the legislators’ actions are likely to be heavily influenced by public interests.

The political spectra of the legislators thus constructed are shown in Figures 2A and 2B for the U.S., and in Figures 2C and 2D for Korea. The spectra of the Twitter users (the average of the positions of the legislators that they follow) are shown as the solid lines in Figures 2A and 2C.

A close examination of these spectra can reveals the disparities between the online and the offline. First, we find a stark indication of the significance of the disparity between Twitter and the offline in the majority-minority reversal in both countries: The majority of Twitter users align themselves with the minority parties. The mean and the median positions, and the overall distribution of the spectra are



**Figure 2.** The political spectra determined from Twitter and roll call data. (A) The Twitter-based spectrum of the legislators, and the spectrum of Twitter users in the US; (B) The roll call-based spectrum of the legislators in the US; (C) The Twitter-based spectrum of the legislators, and the spectrum of Twitter users in Korea; (D) The roll call-based spectrum of the legislators in South Korea. The legislators are indicated by the shapes (one shape corresponds to one legislator) and the spectra of the Twitter users are indicated by solid lines. The party-line splits of the legislators are evident in both the Twitter- and the roll call-based spectra for both countries. The general Twitter public (solid lines in Figures A and C) show a majority-minority reversal, more people aligning themselves with the minority parties.

shown in Figure 2A for U.S. and Figure 2C for Korea. We find that in the US, Republican Senators John McCain (the presidential candidate in 2008) and Jim DeMint (of the populist Tea Party movement) boasting a disproportionately large number of followers are primarily responsible for this behavior. In Korea, the DP (green) overwhelms the GNP (blue) in the Twitter sphere despite being outnumbered by a nearly two-to-one margin in the number of seats in the Assembly.

Second, the comparison between the roll call-based spectra of the legislators as the proxy for the offline public at large and the spectra of the Twitter users also reveals a similar disparity. Although we cannot directly compare the online and offline spectra on the same plot because of the differences in scale, the comparison of relative biases in the distribution is still valid. For the U.S., the spectrum of the Twitter users is heavily skewed towards GOP in sharp contrast to the DEMs who are favored in the spectrum of the offline public (Figures 2A and 2B). The trend is similar in the Korean spectra: the offline public favors GNP in contrast to the Twitter user who favor DP (Figures 2C and 2D).

Third, there also exists a discrepancy between the real political positions of the legislators and their positions constructed from the Twitter data. The party-line split in the positions of the legislators (indicated as solid shapes in the figures) is immediately noticeable for both nations, and accordingly the roll call-based spectra and the Twitter-based spectra of the legislators exhibit some degree of overall agreement (similarity) as measured by Kendall’s  $\tau$ -coefficient (1 means identity; 0 means independence), with  $0.68 \pm 0.02$  for US and  $0.48 \pm 0.01$  for Korea (see Methods section for more detail). When we consider each party separately, however, we find a weaker agreement – for the US parties, the  $\tau$  is  $0.463 \pm 0.003$  for GOP and  $0.291 \pm 0.022$  for DEM. For the Korean parties that is  $0.029 \pm 0.005$  for GNP and  $0.062 \pm 0.051$  for DP.

This indicates that Korean Twitter population is more noticeably impervious to the political positions of the legislators, resulting in the very low level of intra-party correlations between the Twitter based spectrum and the roll call based spectrum of the legislators. In US the correlation is more sizable.

These disparities between the Twitter users and the offline public show the limits of Twitter as a source for fair and accurate opinion mining, especially pertaining to political or social issues. In fact, the majority-minority reversal shows us that Twitter may be functioning as a de facto “alternative media.”

## Quantifying partisanship of twitter users

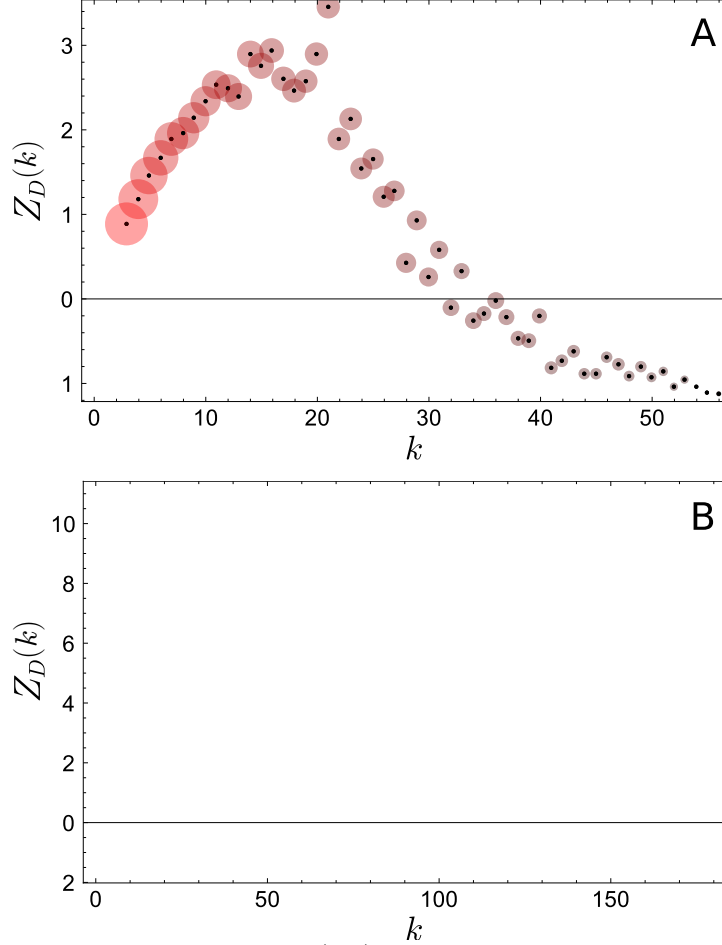
The spread structures of the Twitter users’ spectra in Figures 2A and 2C imply that Twitter users are biased (show partisanship) in choosing which legislators to follow. We can quantify the strength of a user’s bias by measuring the probability that a user follows certain parties, and comparing it with the random probability that a user becomes connected to a party if the connections were made at random. If a user is unbiased they must be identical, and deviations between the two probabilities signify the strength of the bias.

A widely-used measure for the discrepancy between two probability distributions is the Kullback-Leibler (KL) divergence  $D$ . It is non-negative, i.e.  $D \geq 0$ ; 0 when two distributions being compared are identical, and positive otherwise. To account for random fluctuations we calculate the  $Z$ -score of KL divergences (see Methods section for detail)  $Z_D$  as a function of the number of politicians (degree)  $k$  that a user follows, shown in Figure 3.  $Z_D \gtrsim 1$  implies partisanship, while  $Z_D \lesssim -1$  implies the opposite. The radius of the disc around each plot point is proportional to the the log of the number of users with the given degree  $k$ . The figure shows the widespread nature of partisanship on Twitter [9], with the exception of high-degree Twitter users who are more balanced in following each party (see Methods section for more detail).

## Discussion

In this paper we proposed a method to compare the political spectra of the online and offline public. We found significant discrepancies between the online and the offline, demanding caution when one tries to use online social media such as Twitter as source for fair and accurate opinion mining. A quantitative measurement of political biases found widespread partisanship among Twitter users.

While our work shows the potential limits of the correctness of online social media-based findings, it is still meaningful that were able to obtain the political spectra of Twitter users based on their behavior observed objectively. This enables us to overcome uncertainty in methods based on self-assessment of political positions [18–20] caused by the lack of objective, common measure of political positions agreed upon by people. In our method, however, a person’s political position is determined from the objective behavior of all other people in the data, allowing us to construct global and objective political spectra.



**Figure 3.** The Z-score of the Kullback-Leibler (KL) divergence as a function of the number of legislators  $k$  that a Twitter user follows (A) in the US, and (B) in Korea. The radius of each circle is proportional to the log of the number of Twitter users with given  $k$ .  $Z_D \gtrsim 1$  implies a statistically significant level of partisanship (political bias). We see that partisanship is widespread in both countries, except for high-degree Twitter users for whom  $D \rightarrow 0$ .

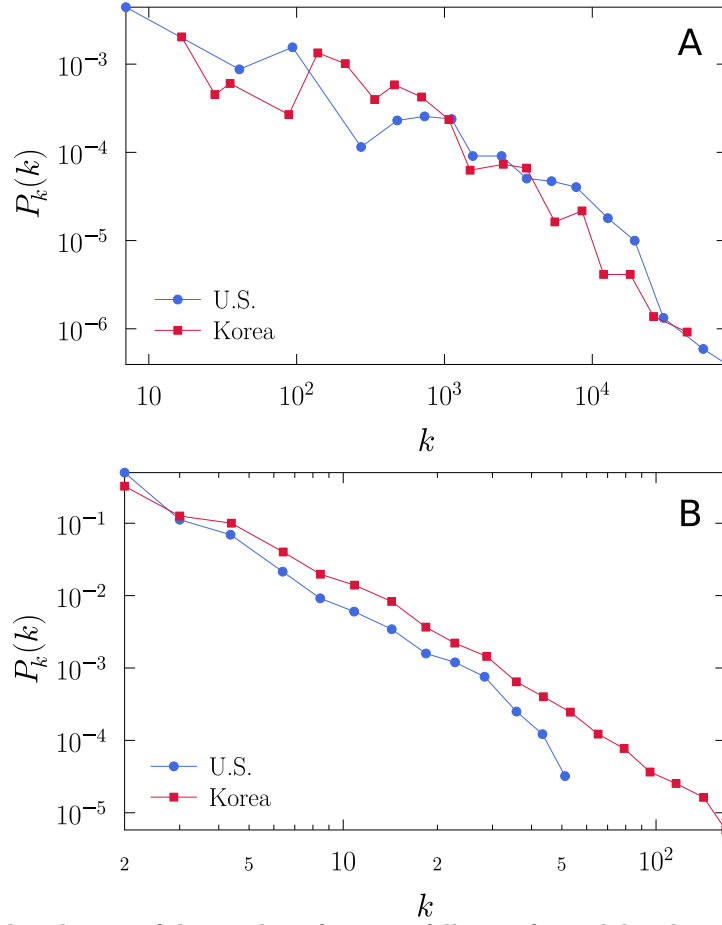
## Methods

### Classical multi-Dimensional scaling

CMDS is a method to embed a set of objects in a pre-defined  $N$ -dimensional space such that the distance between objects match given dissimilarities as much as possible. Specifically, CMDS aims to produce the coordinates of the objects that minimize a loss function called STRAIN [15]. In this paper we set  $N = 1$  to produce a linear political spectrum that agrees with the common way of thinking of political polarization (e.g. left wing versus right wing).

CMDS is performed by the eigendecomposition of the matrix made from the dissimilarities. The quality of the resulting coordinates in reproducing the given dissimilarities is defined as the fraction of the first eigenvalue over the sum of all positive eigenvalues. The values are 0.297 for Korea and 0.872 for US with the roll call data, and 0.113 for Korea and 0.093 for US with the Twitter data. The noticeably high value for the US roll call data may be due to the stable two-party system of the US.

While one could theoretically gain more information from higher-dimension CMDS, we could not observe meaningful patterns allowing straightforward political interpretation beyond the first. Our



**Figure 4.** (A) The distribution of the number of Twitter followers for each legislator. (B) The distribution of the number of legislators that Twitter users follow. All distributions are heavy-tailed.

1-dimensional solutions show clear separations of the legislators along party lines, thus justifying the political interpretation given.

We used the Kulczynski dissimilarity [17]. In Twitter data it is given between two legislators  $i$  and  $j$  as

$$\delta_T(i, j) = 1 - \frac{1}{2} \left| \frac{F_i \cap F_j}{|F_i|} + \frac{F_i \cap F_j}{|F_j|} \right|, \quad (1)$$

where  $F_x$  denotes the set of the followers of legislator  $x$ . The Kulczynski dissimilarity is recommended as an alternative to the commonly used Jaccard measure when the size of  $F_x$  exhibits a wide range as in our Twitter data (Figure 4) [21].

The dissimilarity between legislators  $i$  and  $j$  in the roll call data is, similarly,

$$\delta_R(i, j) = 1 - \frac{1}{2} \left| \frac{V_i \cap V_j}{|V_i|} + \frac{V_i \cap V_j}{|V_j|} \right|, \quad (2)$$

where  $V_x$  denotes the set of *nay* (opposition to a bill's passing) votes of legislator  $x$ . We ignore the *yea* votes here. The motivation for this is that most bills that pass are bipartisan, potentially limiting the discriminating power of the *yea* votes.

## Notes on the spectrum of Twitter users

From the legislator-Twitter data we estimated the spectrum of the legislators first on their Twitter followership, and then used it to obtain the spectrum of Twitter users. A dual approach could have been employed where we perform CMDS on the Twitter users by defining the dissimilarities between the users based on the set of legislators that they follow, and determining the positions of the legislators based on the results. While there is no reason to object this approach in principle, its computational cost was prohibitive, requiring several tens of gigabytes of memory to store the dissimilarity matrix alone.

### Kendall's $\tau$ rank correlation coefficient

Kendall's rank correlation  $\tau$  quantifies the similarity between two orderings of objects. Letting  $(x_1, x_2, \dots, x_n)$  and  $(y_1, y_2, \dots, y_n)$  be the orderings we wish to compare, Kendall's  $\tau$  rank correlation coefficient is defined as

$$\tau = \frac{2(C - D)}{n(n - 1)}. \quad (3)$$

In this definition  $C$  is the number of  $(i, j)$  pairs such that the  $(i, j)$  satisfies  $x_i > x_j$  and  $y_i > y_j$  or it satisfies  $x_i < x_j$  and  $y_i < y_j$ .  $D$  is the number of  $(i, j)$  pairs such that the  $(i, j)$  satisfies  $x_i > x_j$  and  $y_i < y_j$  or it satisfies  $x_i < x_j$  and  $y_i > y_j$ .

Errors in Kendall's  $\tau$  were estimated through the jackknife method [22].

### Z-score of the Kullback-Leibler divergence

Kullback-Leibler divergence (KL divergence) measures deviation of a probability distribution from a reference probability distribution. Here the reference probability distribution is the fractions of seats occupied by each party, i.e. the probability that a random user would follow a legislator from a given party.

Let  $P_i$  be the fraction of the number of legislators from party  $i$  among the number of the legislators, and  $p_{ij}$  be the fraction of the number of legislators from party  $i$  followed by Twitter user  $j$  among all legislators followed by  $j$ . Then the KL divergence  $D_j$  of Twitter user  $j$  is

$$D_j = \sum_{i=1}^N p_{ij} \log \frac{p_{ij}}{P_i}, \quad (4)$$

where  $N$  is the number of parties. KL divergence is non-negative. A large  $D_j$  means that user  $j$ 's followership pattern does not agree with the reference probability, i.e. favors one party disproportionately when in a two-party system.

While the average KL divergence  $\langle D \rangle_k$  of Twitter users who follow  $k$  legislators can be obtained easily from the empirical data, for statistical significance we use the  $Z$ -score

$$Z_D(k) = \frac{\langle D \rangle_k - \mu_k}{\sigma_k}, \quad (5)$$

where  $\mu_k$  and  $\sigma_k$  are the expectation and standard deviation of the KL divergence when a user choose to follow a party according to the reference probability  $P_i$ . If we assume a Twitter user follows  $k$  legislators randomly with uniform distribution, then the probability of  $\vec{n} = (n_1, \dots, n_N)$ , the vector of the number of legislators in each party that the user follows, is given by the multinomial distribution  $f(\vec{n}; \vec{P})$  where  $n_1 + \dots + n_N = k$  and  $\vec{P} = (P_1, \dots, P_N)$ . Then we have

$$\mu_k = \sum_{n_1 + \dots + n_N = k} f(\vec{n}; \vec{P}) \sum_{i=1}^N \frac{n_i}{k} \log \frac{n_i}{kP_i} \quad (6)$$



and

$$\sigma_k^2 = \sum_{n_1 + \dots + n_N = k} f(\vec{n}; \vec{P}) \left( \sum_{i=1}^N \frac{n_i}{k} \log \frac{n_i}{kP_i} \right)^2 - \mu_k^2. \quad (7)$$

$Z_D(k) \gtrsim 1$  indicates that Twitter users follow a biased set of legislators, and that the bias is larger than the typical random fluctuation. Similarly,  $Z_D(k) \lesssim -1$  means that Twitter users follow the parties in close agreement with the reference probability  $P_i$ . As  $k$  approaches the actual number of legislators  $p_{i,j}$  converges to  $P_i$ ,  $Z_D(k)$  becomes more negative (since  $D \rightarrow 0$ ).

## Supporting Information

### S1 Dataset

**List of the Korean legislators.** Each line is a record for a legislator. The first column is the Twitter ID, and the second column is the name. The last column is the party of the legislator. The file is encoded with UTF-8 for Korean characters.

### S2 Dataset

**Followership data for the Korean legislators.** Each line is a followee-follower relation. The first column is the Twitter ID of a legislator, and the second column is the Twitter ID of a follower of the legislator.

### S3 Dataset

**Roll call votes data of the Korean legislators.** Each line is a record of a legislator. The first column is the name of the legislator, and the rest columns are votes. A vote is encoded as 1 for yea, 2 for nay, 3 for abstention, and 4 for absence. The file is encoded with UTF-8 for Korean characters.

### S4 Dataset

**List of the US legislators.** The format is the same with S1 Dataset.

### S5 Dataset

**Followership data for the US legislators.** The format is the same with S2 Dataset.

### S6 Dataset

**Roll call votes data of the US legislators.** The format is the same with S3 Dataset.

## Acknowledgments

We thank Z. Hyung, C. Han, and B. Kahng.

## References

1. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, Brewer D, et al. SOCIAL SCIENCE: Computational Social Science. Science. 2009 Feb;323(5915):721–723.
2. Giles J. Computational social science: Making the links. Nature. 2012 Aug;488(7412):448–450.

3. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? In: WWW '10: Proceedings of the 19th international conference on World wide web. New York, NY, USA: ACM; 2010. p. 591–600.
4. Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in twitter: The million follower fallacy. In: 4th International AAAI Conference on Weblogs and Social Media (ICWSM); 2010. .
5. Golder SA, Macy MW. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*. 2011 Sep;333(6051):1878–1881.
6. Conover M, Ratkiewicz J, Francisco M, Gonçalves B. Political Polarization on Twitter. In: Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM); 2011. p. 89–96.
7. Duggan M, Smith A. Social media update 2013. Pew Research Center; 2014.
8. Hargittai E. Whose Space? Differences Among Users and Non-Users of Social Network Sites. *J Comput-Mediat Comm*. 2007 Oct;13(1):276–297.
9. Adamic LA, Glance N. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In: Proceedings of the 3rd International Workshop on Link Discovery. LinkKDD '05. New York, NY, USA: ACM; 2005. p. 36–43.
10. Parmelee JH, Bichard SL. Politics and the Twitter revolution: how tweets influence the relationship between political leaders and the public. Lanham, Md.: Lexington Books; 2012.
11. Chang D. Korean Politics on Twitter : Networks of Politicians and Voters. *Journal of communication research*. 2011;48:80–107.
12. ;. The two Korean parties changed their names recently. GNP was replaced by New Frontier Party, and DP was replaced by New Politics Alliance for Democracy.
13. Poole KT, Rosenthal H. The polarization of American politics. *J Polit*. 1984;46(04):1061–1079.
14. Porter MA, Mucha PJ, Newman MEJ, Warmbrand CM. A network analysis of committees in the US House of Representatives. *Proc Natl Acad Sci USA*. 2005;102(20):7057–7062.
15. Cox TF. Multidimensional scaling. 2nd ed. No. 88 in Monographs on statistics and applied probability. Boca Raton: Chapman & Hall/CRC; 2001.
16. Brazill TJ, Grofman B. Factor analysis versus multi-dimensional scaling: binary choice roll-call voting and the US Supreme Court. *Soc Networks*. 2002;24(3):201–229.
17. Hubálek Z. Coefficient of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol Rev*. 1982 Nov;57(4):669–689.
18. Greenwald AG, Banaji MR. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol Rev*. 1995 Jan;102(1):4–27.
19. Burdein I, Lodge M, Taber C. Experiments on the Automaticity of Political Beliefs and Attitudes. *Polit Psychol*. 2006 Jun;27(3):359–371.
20. Nosek BA, Grahamm J, Hawkins CB. Implicit political cognition. In: Gawronski B, Payne BK, editors. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. The Guilford Press; 2010. p. 548–564.
21. Hausdorf B, Hennig C. Biotic element analysis in biogeography. *Syst Biol*. 2003;52(5):717–723.
22. Newman MEJ. Monte Carlo methods in statistical physics. Oxford : New York: Clarendon Press ; Oxford University Press; 1999.